Journal of Computer and Knowledge Engineering, Vol. 7, No. 2, 2024. (51-64)

Ferdowsi University of Mashhad

Journal of Computer and Knowledge Engineering

https://cke.um.ac.ir

Information and Communication Technology Association of Iran

Machine Learning Classifiers and Data Synthesis Techniques to Tackle with Highly Imbalanced COVID-19 Data

Research Article

Avaz Naghipour[1], Mohammad Reza Abbaszadeh Bavil Soflaei[2], Mostafa Ghaderi-Zefrehei[3]

DOI: 10.22067/cke.2024.88940.1121

Abstract The COVID-19 pandemic has highlighted the urgent need for rapid and accurate diagnostic methods. In this study, we evaluate three machine learning models—Random Forest (RF), Logistic Regression (LR) and Decision Tree (DT)—for detecting COVID-19 trained on preprocessed imbalanced datasets. The dataset used in this study is heavily imbalanced, with 5086 negative and 558 positive cases, posing a significant challenge for effective model training. To this end, we demonstrate the capability of two advanced data synthesis algorithms, Conditional Tabular Generative Adversarial Network (CTGAN) and Tabular Variational Autoencoder (TVAE), in addressing the class imbalance inherent in the dataset. The classifiers trained on the original as well as the balanced datasets were evaluated for comparison. Our findings reveal that RF obtains the highest accuracy of 98.83% on the CTGAN-balanced dataset. In conclusion, our results verify the potential of coupling data synthesis with traditional machine learning for the diagnosis of COVID-19. We hope that we become a valuable contributor to the ongoing AI for pandemic.

Keywords *COVID-19 Detection, Machine Learning, CTGAN, TVAE, Class Imbalance.*

1. Introduction

In late 2019, a pneumonia outbreak originated in Wuhan, China, which was subsequently identified as being caused by the SARS-CoV-2 virus by the World Health Organization (WHO) [1]. SARS-CoV-2 is an enveloped virus with a positive-sense, single-stranded RNA genome [2]. This virus primarily targets the human respiratory system and is highly transmissible through respiratory droplets from coughing, sneezing, and direct physical contact [3]. Additionally, it can spread via contact with contaminated surfaces, where the virus can persist for several days depending on environmental conditions [4]. Common symptoms of the infection include fever, dry cough, loss of taste and smell, sore throat, and muscle pain [2]. The pandemic has had widespread impacts, leading to the postponement of school examinations, closure of offices, and widespread layoffs [5]. Estimates from the World Health Organization (WHO) show that the full death toll associated directly or indirectly with the COVID-19 pandemic between 1 January 2020 and 31 December 2021 was approximately 14.9 million [6]. The recent surge in data science has empowered healthcare professionals by providing tools to analyze massive datasets of health information for disease detection. This progress is driven by various techniques like deep learning, data mining, and especially machine learning (ML). However, a key challenge remains: selecting the most appropriate ML algorithms that can learn effectively from existing data and make accurate predictions for entirely new cases [2]. ML is crucial in the healthcare sector, particularly for diagnosing diseases, detecting outbreaks, and preventing illnesses. ML algorithms are employed for numerous purposes, including predicting diabetes [7], forecasting the progression of Alzheimer's disease [8], heart disease [9], and other medical conditions. Due to the scarcity of tabular data on COVID-19, we tested our hypothesis using a dataset available on Kaggle (at this link), which clearly represents the clinical symptoms of COVID-19. This dataset, like many others in the field of ML, is heavily imbalanced, containing 5086 negative cases and 558 positive cases, resulting in a 1:9 ratio. Training ML models on imbalanced datasets poses several challenges: the models tend to be biased towards the majority class, leading to poor performance in detecting the minority class [10]. This imbalance can result in lower recall for the minority class, skewed accuracy metrics, and an overall decrease in the model's ability to generalize well to new, unseen data [11]. Our study seeks to improve the classification of COVID-19 by conducting a thorough

* Manuscript received: 2024 July 15, Revised, 2024 October 4, Accepted, 2024 November 6.

[1] Corresponding Author. Assistant Professor, Department of Computer Engineering, University College of Nabi Akram, Tabriz, Iran. **Email:** naghipour@ucna.ac.ir.

[2] Master of Artificial Intelligence, Department of Computer Engineering, University College of Nabi Akram, Tabriz, Iran.

[3] Associate Professor, Department of Animal Science, University of Yasouj, Yasouj, Iran.

comparative analysis of three machine learning (ML) models: Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT). Additionally, to mitigate the effects of imbalanced dataset, we investigate two advanced data augmentation techniques: Conditional Tabular Generative Adversarial Network (CTGAN) [12] and Tabular Variational Autoencoder (TVAE) [12], to balance the dataset. To boost classification accuracy and evaluate the effectiveness of different oversampling techniques, we trained each ML model on data balanced using these techniques. This allowed us to compare the models' performance based on the quality of the synthetic data generated by each technique. Additionally, we ensured the effectiveness of data balancing by comparing the performance of these models against models trained on both the original unbalanced dataset and a baseline model. Our results show exceptional accuracy and performance metrics, exceeding those reported in other studies.

The structure of the study is outlined as follows: Section 2 provides an overview of the relevant literature related to our study. In Section 3, the proposed methodology is detailed, with a thorough explanation of its components and techniques like CTGAN and other used methods. Section 4 presents the experimental findings, including the results of the proposed method. Section 5 is the discussion section where the findings of the study are analyzed and compared with other studies, along with an evaluation of the proposed model’s strengths and limitations. Finally, the article is concluded in Section 6, with a summarization of the key findings and discussion about implications for future research in the field.

2. Literature Review

Machine learning has been extensively employed in various domains to address challenges posed by the COVID-19 pandemic. This section reviews existing literature in the fields of COVID-19 vaccine uptake prediction, COVID-19 detection through medical images, and techniques to address class imbalance in datasets. These studies are grouped based on the methodologies they use, highlighting the relevance of each to our research, and comparing the datasets and performance metrics where appropriate.

a. Machine Learning for COVID-19 Vaccine Acceptance and Uptake Prediction

Within the realm of machine learning applied to vaccination studies, recent research has explored predicting vaccine acceptance and identifying key factors influencing uptake. For instance, a study [13] investigates barriers to COVID-19 vaccine uptake in Ghana using a cross-sectional survey and machine learning algorithms. The study identifies significant factors, such as the type of medical facility visited and the presence of underlying conditions, with the random forest model emerging as the most effective predictor. Similarly, another study [14] applied machine learning algorithms to assess COVID-19 vaccine acceptance in countries where residents had already been vaccinated. This study differs from [13] by focusing on vaccine acceptance in different regions and contexts, showcasing machine learning’s versatility in vaccine-related studies. Further, a study [15] applied machine learning algorithms to analyze vaccination rates across states in the United States. While both [14] and [15] utilize machine learning, [14] targets acceptance while [15] investigates actual vaccination rates, highlighting the diverse applications of machine learning in vaccine-related health studies. However, none of these studies address the issue of class imbalance in datasets, which is a crucial aspect of improving predictive accuracy, especially in health-related research. Our study seeks to extend this work by incorporating advanced data balancing techniques like CTGAN and TVAE to tackle this imbalance.

b. Machine Learning and Deep Learning for COVID-19 Detection Using X-ray Images

A significant body of work has focused on using machine learning and deep learning techniques to detect COVID-19 through medical images, particularly X-rays. One such study [2] employed machine learning algorithms to detect lung changes associated with COVID-19 from X-ray images. The models classified X-ray images into categories such as COVID-19 patients, pneumonia patients, and healthy individuals. Among the models tested, VGG-19 with augmentation achieved the best performance, with 99% training accuracy and 98% testing accuracy. This approach shows great potential for enhancing patient prognosis tracking and supporting treatment efficacy studies. Several other studies have also demonstrated the efficacy of deep learning techniques in detecting COVID-19. For instance, a study [16] utilized Convolutional Neural Networks (CNNs) combined with a filter family and the weight-sharing feature extractor SqueezeNet. The study achieved high detection rates using deep learning for COVID-19 cases, illustrating the power of CNN features and neural network classifiers. Similarly, [17] and [18] applied CNN models for COVID-19 detection, achieving accuracies of 90% and 97%, respectively. In particular, [18] proposed a CNN model for detecting COVID-19-associated changes from X-ray images and demonstrated high accuracy across various classes. In another study [19], the authors employed several deep learning models, including CNNs, long short-term memory (LSTM) networks, GANs, and residual neural networks (ResNets), for classifying COVID-19 from other pneumonia causes using chest X-ray images. Among these, the CNN-based approach showed the highest accuracy of 99%. Despite their promising results, these studies predominantly focus on deep learning methods without addressing class imbalance, which can skew results, especially in smaller datasets with uneven class distributions. In contrast, our study applies data balancing techniques such as CTGAN and TVAE, which allow us to effectively balance the dataset and improve the robustness of machine learning models.

c. Handling Class Imbalance in COVID-19 Datasets

Class imbalance is a recurring issue in medical datasets, particularly in COVID-19 studies, as the number of positive cases is often significantly lower than the number of negative cases. Addressing this issue is crucial for improving model performance and ensuring that predictions are not biased toward the majority class. A notable study [5] tackled this challenge by employing

various oversampling techniques, including the Synthetic Minority Oversampling Technique (SMOTE). SMOTE generates synthetic samples of the minority class to balance the dataset, and this study further enhanced it by introducing a modified version called Outlier-SMOTE, which focuses on data points that are farther from others. The proposed method improved performance across several benchmark datasets, including a COVID-19 dataset. While SMOTE and its variants are widely used, they may not always be sufficient to handle more complex data distributions, particularly in tabular data. Our study builds on this by implementing advanced generative approaches such as Conditional Tabular Generative Adversarial Networks (CTGAN) and Tabular Variational Autoencoders (TVAE) to synthesize realistic samples from the minority class. These techniques have shown superior performance in balancing datasets, and our results indicate that CTGAN, in particular, outperforms traditional methods like SMOTE. For instance, our CTGAN-balanced dataset, trained with the Random Forest model, achieved accuracy levels comparable to deep learning models, underscoring the effectiveness of GAN-based approaches in handling class imbalance.

d. Machine Learning for Predicting COVID-19 in Vulnerable Populations and Other Domains

In addition to detection and vaccination prediction, machine learning has been applied to address the unique challenges posed by COVID-19 in vulnerable populations and other sectors. One study [20] evaluated a machine learning model's ability to predict COVID-19 diagnosis among individuals with intellectual and developmental disabilities (IDD). The random forest model, trained on over 700 variables from three major IDD-specific datasets, achieved an accuracy of 62.5%. This demonstrates the challenges in applying machine learning to vulnerable populations where data availability and quality may be limited. Furthermore, machine learning has been applied beyond healthcare. For example, [21] developed a predictive model using routine clinical laboratory test data to forecast patient survival outcomes. The combination of Lasso and SVM algorithms produced an ROC curve area of 0.9277 using just eight clinical parameters. In a non-health-related study [22], machine learning was applied to a global aviation dataset to predict financial distress. This study highlighted the potential of machine learning to provide accurate predictions in industries heavily impacted by the pandemic.

e. Deep Learning for Automatic COVID-19 Diagnosis Using Chest X-rays

Several studies have employed deep learning models to automatically diagnose COVID-19 from chest X-rays. A study [23] modified deep learning architectures such as VGG16, VGG19, ResNet50, and InceptionV3 to classify COVID-19 cases. These models, collectively termed "COV-DLS," achieved high classification accuracies, with Modified-VGG16 achieving the highest at 98.61%. Another study [24] applied transfer learning to automatically detect COVID-19 from chest X-ray images, with the VGG16 model achieving 98% testing accuracy. Finally, another deep learning study [25] explored pre-trained CNN models for the automatic diagnosis of COVID-19 from chest X-rays, using a dataset of over 1,200 CXR images from COVID-19 patients. The VGG16 model achieved the highest accuracy of 98.28%, demonstrating the potential of CNN-based models for rapid and accurate COVID-19 detection. However, similar to the other deep learning approaches reviewed, these models did not consider class imbalance, which can lead to overfitting in imbalanced datasets. Our study, by contrast, addresses this issue by implementing CTGAN and TVAE techniques, providing a more robust solution for handling class imbalance.

In summary, our study fills a critical gap in the literature by focusing on addressing class imbalance in COVID-19 datasets using advanced data synthesis techniques such as CTGAN and TVAE. By comparing the performance of Random Forest, Logistic Regression, and Decision Tree models, our research offers a comprehensive evaluation of machine learning models trained on both balanced and imbalanced datasets, with results that are comparable to or exceed those of deep learning models.

3. Methodology

This section outlines the methodology employed in our study for COVID-19 detection using the COVID-19 dataset. We began by conducting rigorous data preprocessing procedures to ensure data quality and prepare the dataset for input into machine learning models. Preprocessing steps included feature selection, column dropping, label encoding, train/test splitting, and feature standardization. All experiments were performed using Python 3.9.18 on a system running Windows 11 with 16GB of RAM, an NVIDIA RTX3070TI graphics card, and an AMD Ryzen™ 9 6900HX CPU, ensuring computational efficiency and accuracy in our analyses. Detailed descriptions of each step in our methodology are provided in the following subsections. Figure 1 provides a visual diagram of the proposed ensemble classifier.

a. Dataset Description

The dataset used in this study is the COVID-19 dataset which is a publicly available dataset at this link, and it is an invaluable resource in the realm of healthcare and machine learning. This dataset comprises anonymized data from patients at Hospital Israelita Albert Einstein in São Paulo, Brazil, who underwent SARS-CoV-2 RT-PCR and additional laboratory tests during their hospital visits [5]. It includes 5644 test samples from various patients, evaluated across 111 attributes including Haemoglobin, Platelets, and Arterial Blood Gas Analysis. Among these samples, only 553 individuals tested positive for COVID-19. This reveals a notable class imbalance, with a ratio of 1:9 (minority to majority class).

b. Preprocessing

In this section, we provide a comprehensive overview of the preprocessing techniques employed to prepare the dataset for machine learning analysis. Each step is meticulously detailed, covering approaches for feature selection, label encoding, train/test partitioning, and feature standardization. We also explain the significance

and reasoning behind hyperparameter tuning, using `GridSearchCV` to optimize the performance of our classifiers.

c. Feature *Selection*

Given the heavily imbalanced nature of our dataset, the dataset was also significantly influenced by null values. This required additional feature selection and reduction to ensure the quality of data and performance of our models. Initially, features that had a null value percentage of more than 90% were removed. These features are likely to have very few and they are too insubstantial to contribute any inference into the modeling procedure. It is best to remove them to ensure a higher overall quality of the dataset. Next, other features that were taken out are 'Patient ID', 'Patient admitted to regular ward (1=yes, 0=no)', 'Patient admitted to semi-intensive unit (1=yes, 0=no)' and 'Patient admitted to intensive care unit (1=yes, 0=no)' as these features were not informative and not related to our objective of predicting whether a patient was positive or not. Subsequently, features with zero variance, particularly feature 'Parainfluenza 2', was dropped out. This meticulous feature selection ensured that only the most relevant and informative features were retained for subsequent analysis. Furthermore, according to study [5], we also conducted a similar preprocessing step. The last 19 columns in our dataset were associated with the presence of antigens, represented as binary values (0 or 1). Due to significant null values in these columns, we calculated a row-wise sum. Subsequently, all 19 columns were merged into a single column named ‘other_disease’. It was determined by the authors of [5] that 13% of the patients tested positive for at least one antigen. Any remaining scattered null values were replaced with the mean of their respective non-null values.

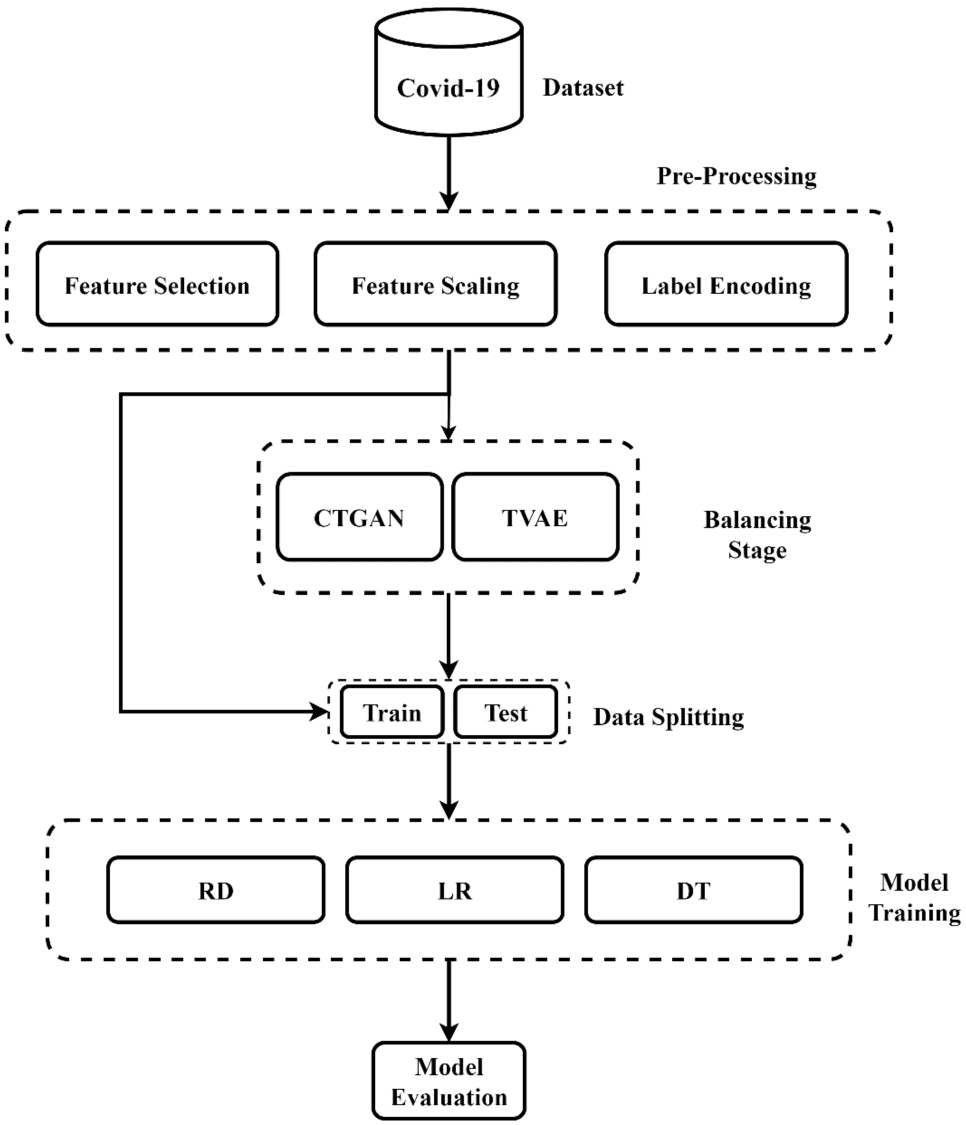


Figure 1. Flowchart of the Proposed Method

3.2.2 Data Splitting

Data splitting is a fundamental technique in ML that involves dividing the dataset into separate subsets for training, validation, and testing. This method facilitates model evaluation and ensures that classifiers are assessed on unseen data, thereby improving their generalizability. In our study, we partitioned the dataset into training and testing sets using the `train_test_split` function from the Python scikit-learn [26] package. Specifically, we adopted an 80:20 split ratio for the classifiers in our proposed method.

3.2.3 Feature Scaling

Feature scaling is a crucial component in machine learning pipelines, ensuring that the range of features is standardized. This standardization reduces the impact of varying magnitudes and enhances the convergence of optimization algorithms. In our study, we utilized the StandardScaler module from the scikit-learn [26] library to perform feature scaling. The standardization process is mathematically expressed as (1)

$$Z = \frac{x - \mu}{\sigma} \quad (1)$$

In this transformation, Z represents the standardized value of the feature, while x denotes its original value. The parameters μ and σ represent the mean and standard deviation of the feature, respectively. Through this mathematical process, the distribution of the feature is effectively centered around a mean of zero with a standard deviation of one.

d. Hyperparameter Tuning

Hyperparameter tuning is a meticulous process aimed at selecting the optimal hyperparameters for a machine learning model, enhancing its functionality and generalizability. This comprehensive search across specified parameter values refines model performance and tailors it to the dataset's nuances. In our research, we employed *GridSearchCV* [26] from the scikit-learn library to determine the best hyperparameters for all the classifiers used. The results of this grid search, detailing the optimal parameters for each model, are summarized in Table 1.

e. Machine Learning Classifiers

Machine learning (ML) algorithms represent a significant advancement over classical algorithms, possessing the ability to learn from data and improve their performance autonomously [27]. By leveraging machine learning techniques, systems can enhance their capabilities and adapt to various environments without explicit programming. ML algorithms are broadly categorized into supervised and unsupervised learning approaches, each offering unique capabilities for data analysis and prediction [28]. The following section introduces the ML algorithms used as classifiers in our methodology, implemented with the scikit-learn [26] Python package, specifically for the detection of COVID-19. The selection of Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT) classifiers was driven by their complementary strengths, simplicity, and ease of implementation. RF is robust and handles imbalanced datasets effectively, providing high accuracy and insights into feature importance. LR offers simplicity and interpretability, serving as a solid baseline for comparison. DTs are not only easily interpretable but also capable of capturing non-linear relationships. Using these classifiers allows us to evaluate the impact of data synthesis techniques like CTGAN and TVAE comprehensively, ensuring a thorough analysis of COVID-19 detection performance.

Random Forest (RF). Random Forest is an ensemble learning technique that involves training multiple decision trees during the training phase [29]. In this method, each tree within the forest independently predicts the target variable, and the final prediction is determined through majority voting among the predictions of all trees. The primary advantage of Random Forests lies in their ability to reduce classification errors compared to traditional classifiers, while also being less prone to overfitting [7].

Table 1. Best Hyperparameters Obtained through GridSearchCV

	Hyperparameters with		
Model Name	**Original Data**	**CTGAN-Balanced**	**TVAE-Balanced**
LR	C:1, penalty:l2, solver:liblinear	C:1, penalty:l2, solver:liblinear	C:1, penalty:l2, solver:liblinear
DT	max_depth:10, min_samples_split:5	max_depth:10, min_samples_split:5	max_depth:5, min_samples_split:2
RF	max_depth:None, min_samples_split:2, n_estimators:100	max_depth:None, min_samples_split:5, n_estimators:200	max_depth:None, min_samples_split:2, n_estimators:100

Decision Trees (DTs) are constructed, and their majority voting is calculated as follows (2) [7]: Let $Cq(p)$ represent the class prediction of the qth Random Forest, and MV denotes the majority vote of the constructed Decision Trees.

This approach combines the predictions of multiple decision trees to produce a final prediction that is generally more robust and accurate than the prediction of any individual tree.

$$c(p) = MV\{C_q(p)\} \tag{2}$$

Logistic Regression (LR). Logistic Regression is a linear classification algorithm that models the probability of a binary outcome using a logistic function. It is employed to estimate the likelihood of an event occurring based on predictor variables [30,31]. The algorithm utilizes the sigmoid function (3), which maps the input to the range between 0 and

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3}$$

Decision Tree (DT). Decision Tree is a versatile and intuitive machine learning algorithm used for both classification and regression tasks. It operates by recursively partitioning the feature space into regions, guided by the values of input features, in a hierarchical manner. At each node of the tree, a decision is made based on a feature's value, aiming to maximize information gain or minimize impurity, such as entropy or Gini impurity. The decision tree algorithm selects the feature and threshold that optimize information gain or minimize impurity at each node, resulting in a hierarchical structure that facilitates interpretation and decision-making. Decision trees are favored for their simplicity, interpretability, and capability to handle both numerical and categorical data. Additionally, ensemble methods like Random Forests and Gradient Boosting Trees extend the capabilities of decision trees, enhancing their predictive power and robustness.

f. Our Approach to Balancing the Dataset

In addressing the class imbalance in our dataset, we employed two advanced data synthesis techniques: Conditional Tabular Generative Adversarial Network (CTGAN) and Tabular Variational Autoencoder (TVAE). Our primary goal was to synthesize additional data to equalize the number of instances between the negative and positive classes. Initially, our dataset comprised 5086 negative and 558 positive cases, creating a significant imbalance that posed a challenge for effective machine learning model training. To mitigate this issue, we adopted the following approach:

I. Training on Positive Cases: We exclusively trained both CTGAN and TVAE on the 558 positive cases. This step was crucial to ensure that the models could accurately learn the data distribution specific to the positive class.
II. Synthesizing Data: After training, both CTGAN and TVAE were utilized to generate synthetic data. The aim was to create enough synthetic positive cases to balance the dataset. Specifically, we synthesized 4528 additional positive cases. This number was calculated to match the number of negative cases (5086), ensuring an equal distribution.
III. Combining Data: The 4528 synthetic positive cases generated by CTGAN and TVAE were then added to the original dataset. This resulted in a balanced dataset with 5086 instances each of negative and positive cases.
IV. By employing this method, we ensured that our machine learning models were trained on a balanced dataset, thereby enhancing their ability to accurately detect COVID-19 cases without bias towards the majority class. This approach demonstrates the effectiveness of CTGAN and TVAE in generating synthetic data to address class imbalances in datasets. Figure 2 demonstrates the class distribution of both balanced and imbalanced versions of the dataset.

g. Utilization of CTGAN for Data Balancing

In numerous real-world scenarios, datasets often exhibit significant class imbalance, where one class greatly outweighs the others [32]. This class imbalance poses challenges for machine learning models, as they typically struggle to perform well on the minority class due to a lack of sufficient examples for learning. CTGAN (Conditional Tabular Generative Adversarial Network) [12] emerges as a powerful solution for generating synthetic data that accurately replicates the distribution of real-world data, leveraging the capabilities of Generative Adversarial Networks (GANs) [33]. CTGAN was specifically developed to address various challenges encountered in generating synthetic tabular data in the field of Informatics in Medicine, including handling mixed data types, non-Gaussian and multi-modal distributions, and highly imbalanced categorical columns. Additionally, the algorithm combines softmax (4) and tanh functions in its output to effectively generate a blend of discrete and continuous columns simultaneously [34].

$$f(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}} \tag{4}$$

Tan$h$ activation function is shown in Eq. (5)

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{5}$$

CTGAN workflow is listed in three steps below [10]:

- **Identification of continuous columns:**

CTGAN initially identifies each continuous column and determines the number of modes present within them. This is achieved by fitting the modes into a Gaussian mixture using the Variational Gaussian Mixture model (VGM) [35].

- **Computing probability density:**

Then, CTGAN computes the probability density for each column value in every row, thereby assessing the likelihood of each value occurring.

- **Sampling and normalization:**

Finally, CTGAN samples a mode based on the calculated

probability density and uses this sampled mode to normalize the value in the new row. This process ensures that the generated data accurately reflects the distribution observed in the original dataset.

Figure 2 visually illustrates both the imbalanced and balanced datasets, vividly depicting the skewed class distribution in the imbalanced scenario. However, after applying CTGAN, the balanced version emerges, ensuring equitable representation of both classes. This balance not only addresses the skewed distribution but also contributes to enhanced model performance by providing sufficient examples for learning from each class. Additionally, Table 2 represents the hyperparameters used in CTGAN.

h. Other Balancing Techniques

In our study, instead of solely relying on CTGAN, we adopted another approach by evaluating a widely adopted data synthesis technique: Tabular Variational Autoencoder (TVAE) [12]. TVAE represents an established method in the field for addressing class imbalance within datasets. The same mentioned classifiers were also trained on data balanced by TVAE. The results of TVAE are discussed in the next sections.

i. Tabular Variational Autoencoder

In addition to evaluating CTGAN, we conducted thorough comparative analysis by testing another widely utilized method in balancing and oversampling techniques, namely Tabular Variational Autoencoder (TVAE). Variational Autoencoders (VAEs) are advanced deep generative models widely used for generating synthetic data from real datasets. VAEs consist of two main components: The Encoder and the Decoder. The Encoder compresses the input data into a latent probability distribution, while the Decoder generates new instances based on this inferred latent space. This approach allows VAEs to learn and recreate the original input data from the learned probability distribution. In recent research, Xu et al. [12] introduced the Tabular Variational Autoencoder (TVAE), a variant of VAE specifically tailored for tabular data. This study applied TVAE to generate synthetic COVID-19 data. The TVAE model takes real COVID data and analyzes feature variables based on their statistical and probabilistic distributions. The sensitivity of synthetically generated data is crucial in biomedical domains, especially in disease research. The TVAE model addresses this sensitivity by incorporating the Evidence Lower Bound Loss (ELBO) [36]. To illustrate the computational process, the TVAE model is represented by the following Eq. (6):

$$G(x) = T(Decoder(Encoder(x))) \quad (6)$$

Here, $G(x)$ represents the generated synthetic data of COVID-19 instances, x represents a real data instance of COVID-19 instances, and T is the function of the tabular variational autoencoder that takes x as input and generates $G(x)$. The $Encoder()$ function learns the latent distribution from real data, while the $Decoder()$ function generates synthetic data by analyzing these latent distributions. The TVAE method operates in a semi-supervised manner for the synthetic generation of COVID-19 data. The model first learns the latent space distribution of real data and then replicates this data while minimizing the loss function. This semi-supervised nature makes the TVAE model well-suited for use in the biomedical domain, particularly for generating synthetic data that accurately reflects the characteristics of real-world datasets. In this study, TVAE was used to produce synthetic samples representing both negative and positive for the COVID-19 dataset.

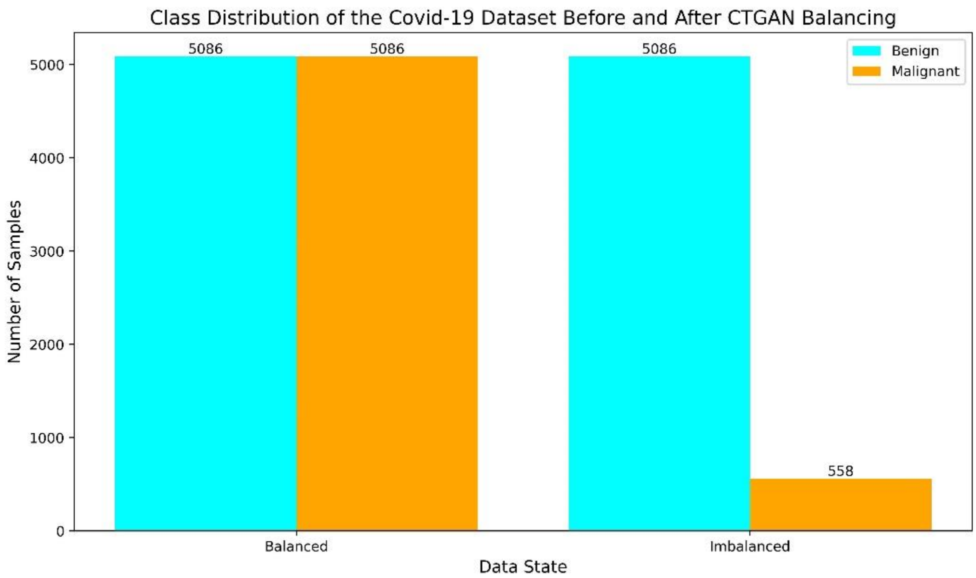


Figure 2. CTGAN-Balanced and Imbalanced versions of the Covid-19 Dataset.

Table 2. Hyperparameters of CTGAN.

Hyperparameters	Values
embedding_dim	128
generator_lr	0.0001
generator_decay	0.00001
discriminator_lr	0.00001
discriminator_decay	0.001
discriminator_steps	3
epochs	500

4. Experimental Results

This section evaluates the effectiveness of the oversampling method using various well-known machine learning metrics, including accuracy, precision, recall, F1-score, and ROC/AUC. Subsection 4.1 provides detailed explanations of these metrics. Additionally, we perform a comparative analysis between the results obtained from the balanced data approach and those from standalone models, highlighting the improvements achieved solely through dataset balancing. The outcomes for the baseline classifiers trained on the original dataset are presented in subsection 4.2. Subsections 4.3 and 4.4 detail the findings for the classifiers trained on CTGAN-balanced data and TVAE-balanced data, respectively.

a. *Performance Metrics*

Before delving into the evaluation metrics employed in this study, it is imperative to understand key terms such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These terms are fundamental in assessing the performance of classification models. In this paper, all performance metrics are explained alongside their corresponding formulas, ensuring clarity and enabling a comprehensive understanding of the evaluation process.

- True Positive (TP): Instances correctly predicted as positive by the model.
- True Negative (TN): Instances correctly predicted as negative by the model.
- False Positive (FP): Instances incorrectly predicted as positive by the model.
- False Negative (FN): Instances incorrectly predicted as negative by the model.

b. Accuracy

Accuracy, a crucial metric in classification tasks, quantifies the proportion of correctly predicted cases, encompassing both true positives and true negatives, out of all instances. The formula for Accuracy is provided in (7)

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \tag{7}$$

c. Precision

Precision, a vital metric in classification assessment, emphasizes the quality of positive predictions by determining the proportion of correctly identified positive instances out of all instances predicted as positive. A higher precision value signifies a lower rate of false positives. The formula for Precision is provided in (8)

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{8}$$

d. Recall

Recall (9), often referred to as sensitivity or true positive rate, assesses the model's capability to identify positive cases accurately. It quantifies the proportion of actual positive instances that the model correctly predicts. A higher recall value indicates a lower rate of false negatives. The formula for Recall is expressed as follows:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{9}$$

e. F1-Score

F1-Score (10), a harmonic mean of precision and recall, provides a balanced evaluation by considering both false positives and false negatives. This metric is especially valuable in scenarios where class distribution is imbalanced. By combining precision and recall, F1-Score offers a comprehensive assessment of the model's performance. The formula for F1-Score is defined as:

$$F1 = 2 \times \frac{Precision \times\ Recall}{Precision + Recall} \tag{10}$$

f. ROC/AUC

The trade-off between true positive rate (recall) and false positive rate is represented graphically by the ROC curve. It provides an overview of the model's overall performance across several thresholds.

g. Baseline Results: Training Results on the Original Dataset

In this subsection, Table 3 presents the results of the three classifiers trained on the original dataset. This table provides a comprehensive analysis of performance metrics, including accuracy, precision, recall, F1-score, and ROC/AUC, for each classifier. It's worth noting that these classifiers were trained on the original, pre-processed COVID-19 dataset. Among these models, LR emerges as the top performer, achieving an accuracy of 88%. Additionally, Figure 3 represents the baseline model accuracies and ROC curves, providing a comprehensive comparison of the models' performance before applying data balancing techniques.

Table 3. Performance of Classifiers, Trained on Original COVID-19 Dataset, with 80:20 Split Ratio.

Model	Label	Precision	Recall	F1-Score	Accuracy	AUC
RF	0	0.90	0.97	0.94	0.88	0.66
	1	0.67	0.35	0.46		
LR	0	0.92	0.96	0.94	0.89	0.72
	1	0.67	0.47	0.55		
DT	0	0.93	0.93	0.93	0.89	0.81
	1	0.59	0.59	0.59		

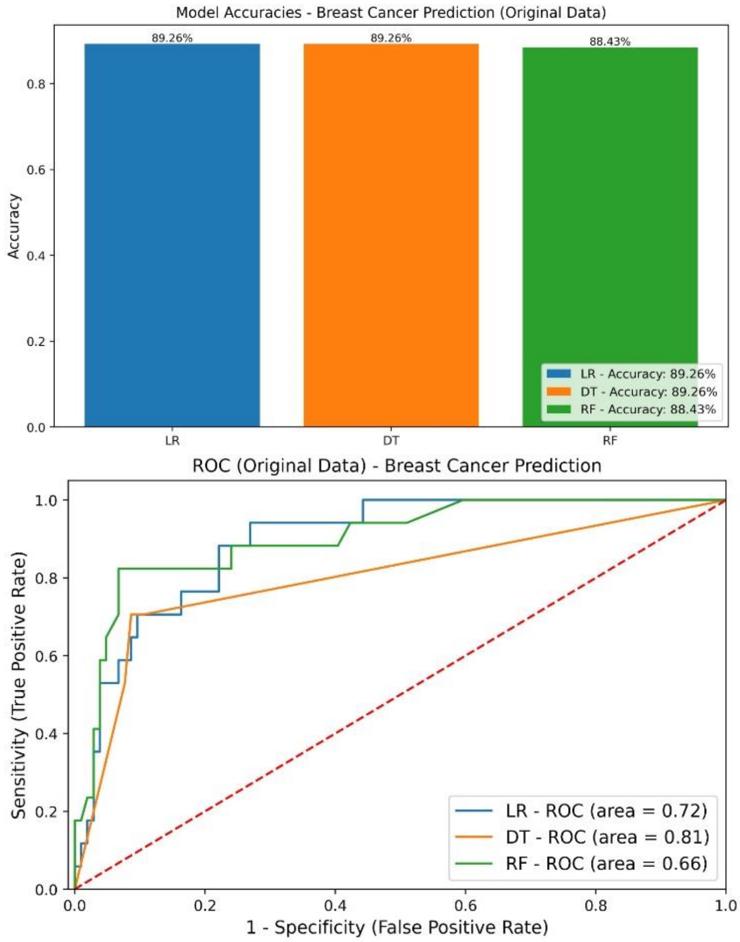


Figure 3. Accuracies and ROC Curves of Models Trained on Original Data.

Table 4. Performance of Classifiers Trained on CTGAN-Balanced Data, with 80:20 Split Ratio.

Model	Label	Precision	Recall	F1-Score	Accuracy	AUC
RF	0	0.93	0.96	0.94	**0.99**	0.97
	1	1.00	0.99	0.99		
LR	0	0.81	0.33	0.47	0.92	0.65
	1	0.93	0.99	0.96		
DT	0	0.91	0.83	0.86	0.97	0.90
	1	0.98	0.99	0.99		

h. CTGAN-Balanced Dataset Results

In this subsection, Table 4 presents the results of training classifiers with CTGAN-balanced data. The classifiers used are Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT). The same performance metrics are calculated and reported to assess the classifiers' effectiveness. As shown in Table 4, RF trained on the CTGAN-balanced dataset is the top performer, achieving an impressive accuracy of 0.9883 and an AUC of 0.97. Additionally, other classifiers also showed improved accuracy by balancing their input data with synthetic data.

i. TVAE-Balanced Dataset Results

In this subsection, the performance of classifiers trained on TVAE-balanced data is presented in Table 5. The same set of metrics is used to evaluate the classifiers' performance. From Tables 5 and 4, it is evident that the RF trained on CTGAN-balanced data outperformed all other RF models trained on both the original data and TVAE-balanced data, achieving an accuracy of 98.83%

.

Table 5. Performance of Classifiers Trained on TVAE-Balanced Data, with 80:20 Split Ratio.

Model	Label	Precision	Recall	F1-Score	Accuracy	AUC
RF	0	0.92	1.00	0.96	0.94	0.90
	1	1.00	0.81	0.90		
LR	0	0.94	0.85	0.89	0.86	0.86
	1	0.74	0.88	0.80		
DT	0	0.94	0.88	0.91	0.88	0.87
	1	0.78	0.88	0.82		

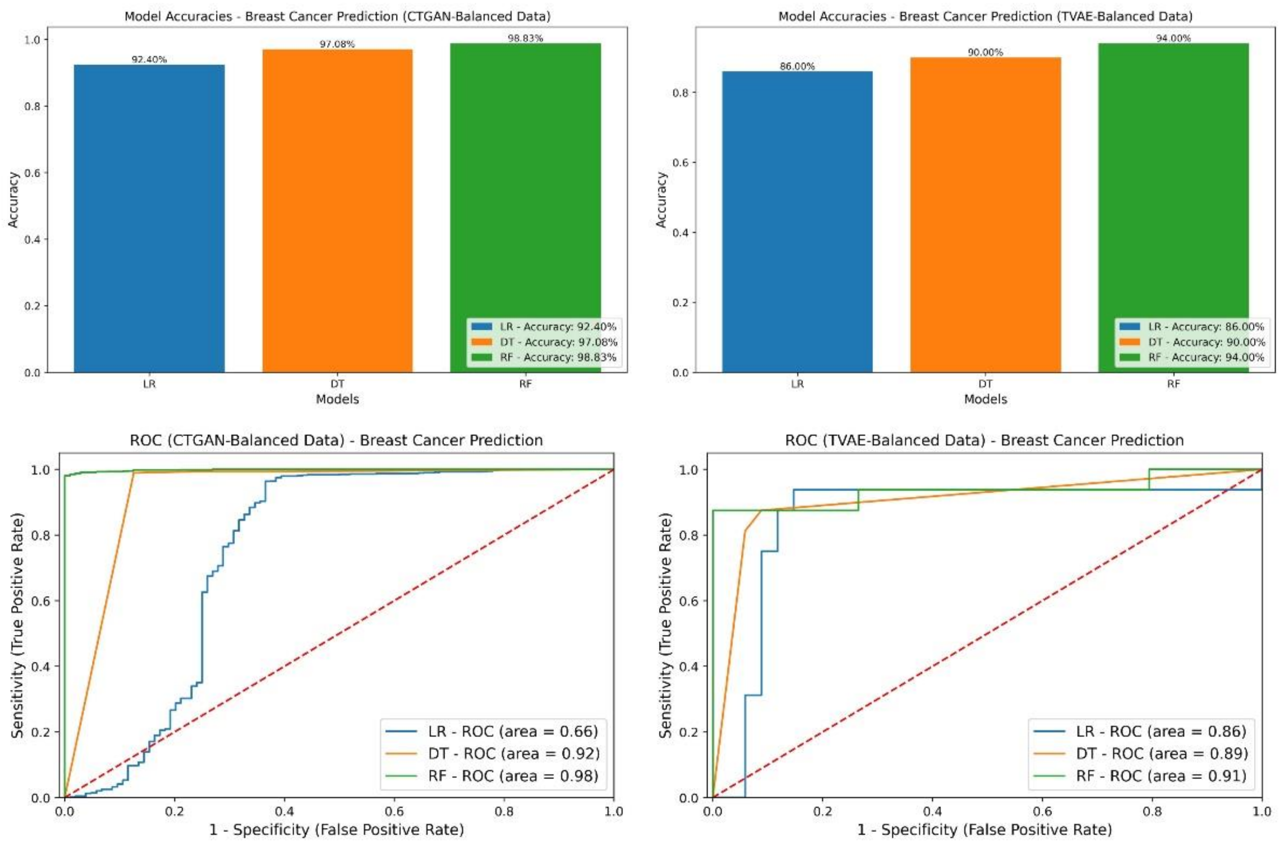


Figure 4. Accuracies and ROC Curves of Models Trained on both CTGAN and TVAE-Balanced Dataset.

Table 6. Comparison with Similar Methods in the Literature

References	Methods	Accuracy
Ahmed et al. [16]	Decision Tree, Random Forest, Neural Network (NN), Naive Bayes, Logistic Regression, and k-nearest neighbor	97.24%
Hemdan et al. [17]	VGG-19	90%
Kumar et al. [23]	Modified-VGG16, Modified-VGG19, Modified-ResNet50, and Modified-InceptionV3	98.61%
Hafeez et al. [18]	CODISC-CNN (CNN based Coronavirus Disease Prediction System for Chest X-rays)	89%
Taresh et al. [25]	Deep transfer learning algorithms including VGG16 and MobileNet	98.28%
Turlapati et al. [5]	Outlier-SMOTE, SMOTE, ADASYN	89-90%
Proposed	Balancing techniques (CTGAN, TVAE) + LR, RF, DT,	**98.83%**

Furthermore, Figure 4 shows the accuracies and ROC curves of models trained on both CTGAN and TVAE-balanced datasets. Notably, the RF classifier trained on the CTGAN-balanced data achieves a higher area under the curve (AUC) compared to the RF classifiers trained with other methods.

5. Discussion

The findings of this study provide significant insights that can enhance the field of COVID-19 detection. By assessing the performance of different machine learning models, specifically Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT), on the Kaggle COVID-19 dataset, we add to the existing body of research and broaden the understanding of machine learning approaches for COVID-19 detection. Our results underscore the effectiveness of data synthesis and oversampling techniques, such as Conditional Tabular Generative Adversarial Network (CTGAN) and Tabular Variational Autoencoder (TVAE), in mitigating dataset imbalances. These insights are highly valuable for healthcare professionals and researchers aiming to improve the accuracy and efficiency of COVID-19 detection and diagnosis.

a. Comparative Analysis with other Studies

In this section, a comparative analysis of our study's findings with those reported in previous research are presented in Table 6.

b. Balanced Dataset Effects on Dataset

Imbalanced datasets have a substantial effect on the performance of machine learning techniques. When data is skewed, ML algorithms tend to favor the majority classes, overlooking those with fewer instances. This bias can undermine the overall performance and quality of the ML model [37]. The dataset used in this study (COVID-19) is heavily imbalanced, with a ratio of 1:9, intensifying the associated challenges. In the field of machine learning, addressing class imbalance has led to the development of various data oversampling and under sampling techniques, such as SMOTE [38] (Synthetic Minority Oversampling Technique), RUS [39] (Random Under-Sampling), and ROS [40] (Random Over-Sampling). However, recent research [35] indicates that Conditional Tabular Generative Adversarial Network (CTGAN) [35] outperforms these traditional methods in data synthesis [12,41]. As demonstrated in our study, employing CTGAN-based data oversampling yields superior results. When utilizing the same classifiers—Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT)—trained on the CTGAN-balanced dataset, we observed an overall enhancement of 5% to 10% in the mentioned evaluation metrics. Studies like current research on pandemics like COVID-19 can be instrumental in enhancing future preparedness and containment strategies. By focusing on early detection, surveillance, and monitoring of potential pathogens in both animal and human populations, promoting wildlife and ecosystem protection, implementing biosecurity measures, and strengthening healthcare infrastructure, we can improve our ability to prevent, detect, and respond to future pandemics effectively.

6. Conclusion and Future Work

The COVID-19 pandemic ravaged the globe, and overwhelmed healthcare systems on a mass scale, exacting an enormous number of lives. Prompt and precise diagnostic approaches have been a critical need. In our study, we evaluated three machine learning models—

Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT)—for detecting COVID-19 using preprocessed datasets. The dataset used in this study is heavily imbalanced, with 5086 negative and 558 positive cases, posing a significant challenge for effective model training. Therefore, we demonstrated the capability of two advanced data synthesis algorithms, Conditional Tabular Generative Adversarial Network (CTGAN) and Tabular Variational Autoencoder (TVAE), in addressing the class imbalance inherent in the dataset. The performance of the models trained on both the original and balanced datasets was then compared. Our findings reveal that RF obtains the highest accuracy of 98.83% on the CTGAN-balanced dataset. We believe that exploiting data synthesis along with classical machine learning approaches holds promise for enhancing the accuracy of COVID-19 diagnosis. This approach could be particularly beneficial in resource-limited settings and developing countries. Moving forward, we recommend the adoption of balanced datasets in training high-performance systems to support effective pandemic response.

Scope for future research lies in incorporating multi-modal data sources, such as merging chest X-ray images with demographic information, comorbidities, and laboratory investigations. Prediction of COVID-19 diagnosis with diverse clinical attributes would only further strengthen the diagnostic model. Also, explainable machine learning endeavors should be initiated to understand the black-box AI decision. Investigation of the approach of model explainability techniques and feature importance methods may help interpret the model decisions in a clinical context, ensuring patient safety, privacy, comfort and their rights are held and protected.

Data Availability

The dataset used in this study is the COVID-19 dataset, which is a publicly available dataset at this link: https://www.kaggle.com/datasets/einsteindata4u/COVID19.

7. Funding

This research has not received any external funding.

Conflict of interest: The writers do not have any relevant conflicts of interest to disclose about the subject matter of this work.

Ethical Approval: Not applicable

Consent to Participate: Not applicable.

Consent to Publish: Not applicable.

8. References

[1] *Naming the coronavirus disease (COVID-19) and the virus that causes it*, World Health Organization, 2020.

[2] S. Hamal et al. and R. M. Gibson. (2024, Jun.). *A comparative analysis of machine learning algorithms for detecting COVID-19 using lung X-ray images.* Decision Analytics Journal. [Online]. 11, p. 100460. Available: 10.1016/j.dajour.2024.100460

[3] Debata, B., P. Patnaik, and A. Mishra. (2020, Sep.). *COVID- 19 pandemic! It's impact on people, economy, and environment.* Journal of public affairs. [Online]. 20(4), p. e2372. Available: 10.1002/pa.2372

[4] Baker, C.A. and K.E. Gibson. (2022, Oct.). *Persistence of SARS-CoV-2 on surfaces and relevance to the food industry.* Current Opinion in Food Science. [Online]. 47, p. 100875. Available: 10.1016/j.cofs.2022.100875

[5] Turlapati, V.P.K. and M.R. Prusty. (2020, Dec.). *Outlier-SMOTE: A refined oversampling technique for improved detection of COVID-19.* Intelligence-Based Medicine. [Online]. 3-4, p. 100023. Available: 10.1016/j.ibmed.2020.100023

[6] *14.9 million excess deaths associated with the COVID-19 pandemic in 2020 and 2021*, WHO, 2022.

[7] B. Amma. (2024, Mar.). *En-RfRsK: An ensemble machine learning technique for prognostication of diabetes mellitus.* Egyptian Informatics Journal. [Online]. 25, p. 100441. Available: 10.1016/j.eij.2024.100441

[8] S. El-Sappagh et al. and H. Saleh. (2021, Sep.). *The role of medication data to enhance the prediction of Alzheimer's progression using machine learning.* Computational Intelligence and Neuroscience. [Online]. (1), 2021. Available: 10.1155/2021/8439655

[9] M. A. Islam et al. and S. Jannaty. (2024, Jun.). *Precision Healthcare: A Deep Dive into* Machine *Learning Algorithms and Feature Selection Strategies for Accurate Heart Disease Prediction.* Computers in Biology and Medicine. [Online]. 176, p. 108432. Available: 10.1016/j.compbiomed.2024.108432

[10] M. R. A. B. Soflaei, A. Salehpour and K. Samadzamini. (2024, Apr.). Enhancing *network intrusion detection: a dual-ensemble approach with CTGAN-balanced data and weak classifiers.* The Journal of Supercomputing. [Online]. 80, pp. 16301-16333. Available: 10.1007/s11227-024-06108-7

[11] A. Luque et al. and A. D. L. Heras. (2019, Jul.). *The impact of class imbalance in* classification *performance metrics based on the binary confusion matrix.* Pattern Recognition. [Online]. 91, pp. 216-231. Available: 10.1016/j.patcog.2019.02.023

[12] L. Xu et al. and K. Veeramachaneni, "*Modeling tabular data using conditional gan,"* Advances in neural information processing systems, 2019.

[13] C. C. Dodoo et al. and J. Mensah. (2024, Jun.). *Using machine learning algorithms to predict COVID-19 vaccine uptake: A year after the introduction of COVID-19 vaccines in Ghana.* Vaccine: X. [Online]. 18, p. 100466. Available: 10.1016/j.jvacx.2024.100466

[14] Oyewola, D.O., E.G. Dada, and S. Misra. (2022, Nov.). *Machine learning for optimizing daily COVID-19 vaccine dissemination to combat the pandemic.* Health and Technology. [Online]. 12, pp. 1277-1293. Available: 10.1007/s12553-022-00712-4

[15]S. M. I. Osman and A. Sabit. (2022, Dec.). *Predictors of COVID-19 vaccination rate in USA: A machine learning approach.* Machine Learning with Applications. [Online]. 10, p. 100408. Available: 10.1016/j.mlwa.2022.100408

[16] M. A. Ahmed et al. and M. Hamid Ali, "*Automatic COVID-19 pneumonia diagnosis from x-ray lung image: A Deep Feature and Machine Learning Solution,"* Journal of Physics: Conference Series, 2021, p. 012099.

[17]E. E. D. Hemdan, M. A. Shouman, and M. E. Karar, "*Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images,"* arXiv preprint arXiv:2003.11055, 2020.

[18]U. Hafeez et al. and H. A. Madni. (2022, Feb.). *A CNN based coronavirus disease prediction system for chest X-rays.* Journal of Ambient Intelligence and Humanized Computing. [Online]. 14(10), pp. 13179-13193. Available: 10.1007/s12652-022-03775-3

[19] H. Mohammad-Rahimi et al. and S. Ghafouri-Fard. (2021, Mar.). *Application of machine learning in diagnosis of COVID-19 through X-ray and CT images: a scoping review.* Frontiers in cardiovascular medicine. [Online]. 8, p. 638011. Available: 10.3389/fcvm.2021.638011

[20] M. D. Broda et al. and A. West. (2024, Jul.). *Understanding COVID-19 infection among people with intellectual and developmental disabilities using machine learning.* Disability and Health Journal. [Online]. 17(3), p. 101607. Available: 10.1016/j.dhjo.2024.101607

[21]Y. Fu et al. and S. Wang. (2024, May.). *Using machine learning algorithms based on patient admission laboratory parameters to predict adverse outcomes in COVID-19 patients.* Heliyon. [Online]. 10(9): p. e29981. Available: 10.1016/j.heliyon.2024.e29981

[22] K. Halteh et al. and K. Kumar. (2024, Mar.). *Using machine learning techniques to assess the financial impact of the COVID-19 pandemic on the global aviation industry.* Transportation Research Interdisciplinary Perspectives. [Online]. 24, p. 101043. Available: 10.1016/j.trip.2024.101043

[23] V. Kumar et al. and O. Cheikhrouhou. (2022, Apr.). *COV-DLS: prediction of COVID-19 from X-rays using enhanced deep transfer learning techniques.* Journal of Healthcare Engineering. [Online]. (1), p. 6216273. Available: 10.1155/2022/6216273

[24] A. Chen et al. and J. Chan, "*Detecting Covid-19 in Chest X-Rays using* Transfer *Learning with VGG16*," *CSBio '20: Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics*, Association for Computing Machinery: Bangkok, Thailand, 2020, pp. 93–96.

[25] M. M. Taresh et al. and M. L. Mutar. (2021, May.). *Transfer Learning to Detect COVID-19 Automatically from X-Ray Images Using Convolutional Neural Networks.* International Journal of Biomedical Imaging. [Online]. 2021(1), p. 8828404. Available: 10.1155/2021/8828404

[26] F. Pedregosa. (2011). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research. [Online]. 12, pp. 2825-2830. Available: cir.nii.ac.jp

[27] H. I. Sarker. (2021, Mar.). *Machine Learning: Algorithms, Real-World Applications and Research Directions.* SN Computer Science. [Online]. 2(3), p. 160. Available: 10.1007/s42979-021-00592-x

[28] M. Alloghani et al. and A. J. Aljaaf. (2019, Sep.). *A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science*. *Supervised and Unsupervised Learning for Data Science*. [Online]. pp. 3-21.

[29] G. Louppe, "*Understanding random forests: From theory to practice,"* arXiv preprint arXiv:1407.7502, 2014.

[30] Y. Belsti et al. and H. Teede. (2023, Nov.). *Comparison of machine learning and* conventional *logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash GDM Machine learning model.* International Journal of Medical Informatics. [Online]. 179, p. 105228. Available: 10.1016/j.ijmedinf.2023.105228

[31] P. Rajendra and S. Latifi. (2021, Jan.). *Prediction of diabetes using logistic regression and ensemble techniques.* Computer Methods and Programs in Biomedicine Update. [Online]. 1, p. 100032. Available:10.1016/j.cmpbup.2021.100032

[32] M. M. Chowdhury, R. S. Ayon and M. S. Hossain. (2024, Jun.). *An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset.* Healthcare Analytics. [Online]. 5, p. 100297. Available: 10.1016/j.health.2023.100297

[33] I. Goodfellow et al. and Y. Bengio, "*Generative adversarial nets,"* Advances in neural information processing systems, 2014.

[34] M. S. K. Inan, S. Hossain and M. N. Uddin.(2023, Jan.). *Data augmentation guided breast cancer diagnosis and prognosis using an integrated deep-generative framework based on breast tumor's morphological information.* Informatics in Medicine Unlocked. [Online]. 37, p. 101171. Available: 10.1016/j.imu.2023.101171

[35] S. Bourou et al. and T. Zahariadis. (2021, Sep.). *A Review of Tabular Data Synthesis Using* GANs *on an IDS Dataset*. Information. [Online]. 12(9), p. 375. Available: 10.3390/info12090375.

[36] D. P. Kingma and M. Welling, "*Auto-encoding variational bayes,"* arXiv preprint arXiv:1312.6114, 2013.

[37] V. W. D. Vargas et al. and J. L. V. Barbosa. (2022, Nov.). *Imbalanced data* preprocessing *techniques for machine learning: a systematic mapping study.* Knowledge and Information Systems. [Online]. 65(1), pp. 31-57. Available: 10.1007/s10115-022-01772-8

[38] N. V. Chawla et al. W. P. Kegelmeyer. (2002, Jun.). *SMOTE: synthetic minority over-sampling technique.*

Journal of artificial intelligence research. [Online]. 16, pp. 321-357. Available: 10.1613/jair.953

[39] T. Hasanin, T. Khoshgoftaar, "*The Effects of Random Undersampling with Simulated Class Imbalance for Big Data*," *IEEE International Conference on Information Reuse and Integration (IRI)*, Lake City, UT, USA, 2018, pp. 70-79.

[40] Imbalanced *learn*. Available: imbalanced-learn.org

[41] I. Ashrapov, "Tabular GANs for uneven distribution," arXiv preprint arXiv:2010.00638, 2020.